Actionable Genes, Core Databases, and Locus-Specific Databases

Amélie Pinard, Morgane Miltgen, Arnaud Blanchard, Hélène Mathieu, Jean-Pierre Desvignes, David Salgado, Aurelie J Fabre, Pauline Arnaud, Laura Barre, Martin Krahn, et al.

Actionable Genes, Core Databases, and Locus-Specific Databases

Amélie Pinard,[1] Morgane Miltgen,[1] Arnaud Blanchard,[1] Hélène Mathieu,[1] Jean-Pierre Desvignes,[1] David Salgado,[1] Aurélie Fabre,[1,2] Pauline Arnaud,[3,4,5] Laura Barré,[1] Martin Krahn,[1,2] Philippe Grandval,[1,6] Sylviane Olschwang,[1,2,7,8] Stéphane Zaffran,[1] Catherine Boileau,[3,4,5] Christophe Béroud,[1,2] and Gwenaëlle Collod-Béroud[1]*

[1]Aix Marseille Univ, INSERM, GMGF, Marseille, France; [2]APHM, Hôpital Timone Enfants, Laboratoire de Génétique Moléculaire, Marseille 13385, France; [3]AP-HP, Hôpital Bichat, Centre National de Référence pour le syndrome de Marfan et apparentés, Paris, France; [4]UFR de Médecine, Diderot Paris Université Paris 7, Paris, France; [5]Inserm U1148, Paris, France; [6]AP-HM, Hôpital de la Timone, Gastroentérologie, Marseille, France; [7]Hôpital Clairval, Ramsay Générale de Santé, Marseille, France; [8]Hôpital Européen, Fondation Ambroise Paré, Marseille, France

For the Next Generation Sequencing special issue

ABSTRACT: **Adoption of next-generation sequencing (NGS) in a diagnostic context raises numerous questions with regard to identification and reports of secondary variants (SVs) in actionable genes. To better understand the whys and wherefores of these questioning, it is necessary to understand how they are selected during the filtering process and how their proportion can be estimated. It is likely that SVs are underestimated and that our capacity to label all true SVs can be improved. In this context, Locus-specific databases (LSDBs) can be key by providing a wealth of information and enabling classifying variants. We illustrate this issue by analyzing 318 SVs in 23 actionable genes involved in cancer susceptibility syndromes identified through sequencing of 572 participants selected for a range of atherosclerosis phenotypes. Among these 318 SVs, only 43.4% are reported in Human Gene Mutation Database (HGMD) Professional versus 71.4% in LSDB. In addition, 23.9% of HGMD Professional variants are reported as pathogenic versus 4.8% for LSDB. These data underline the benefits of LSDBs to annotate SVs and minimize overinterpretation of mutations thanks to their efficient curation process and collection of unpublished data.**

KEY WORDS: **secondary variant; actionable genes; LSDB; databases; NGS**

Introduction

Progress in sequencing technologies have led to the rapid adoption of next-generation sequencing (NGS) in a research context to facilitate the identification of disease-causing genes, especially in the field of rare genetic diseases. Based on their successes, these technologies have been transferred to diagnosis. This switch was not transparent but rather has been accompanied by new ethical issues. In fact, patients are addressed for a specific set of symptoms associated to a particular disease spectrum such as a neuromuscular disease. In the course of the Whole-Exome Sequencing (WES) now routinely proposed in many countries, it is frequent to identify potentially harmful mutations in genes unrelated to these symptoms but which may be of importance for patient follow-up. These discoveries have been named "secondary findings," "incidental findings," or "secondary variants (SVs)" and have to be distinguished from "unsolicited findings" that are found in the genes linked to the tested disease [Matthijs et al., 2016]. Depending on national guidelines, it may be mandatory or not to look for these "SVs" (for more information, see dedicated paper in this issue). Because these findings usually target genes involved in a completely different clinical field, such as cancer predisposing genes, the diagnostic laboratory may not be an expert of these genes while the interpretation of results requires a strong expertise. During the last 25 years, Locus-specific databases (LSDBs) have been slowly developed and maintained to ensure optimal quality of data to facilitate data interpretation. Here, we will review the various resources, LSDB and other databases, that could be used to facilitate the interpretation of these findings and discuss assets and drawbacks.

Materials and Methods

Variant list from Johnston et al. (2012) is available as Supp. Table S1 in Johnston et al. (2012) and Matthijs et al. (2016). List of actionable genes is available in Green et al. (2013). Reports in LSDBs for each of the 318 variants were searched for with an in-house designed Perl script for LOVD Beacon (http://mcupak.github.io/beacon-of-beacons/queries.html) and LOVD Share (http://databases.lovd.nl/shared/genes) core databases. Data from each UMD-LSDB (*APC*, *BRCA1*, *BRCA2*, *MEN1*, *MLH1*, *MSH2*, *MSH6*, *MUTYH*, *TP53*, and *VHL* genes) were searched online at http://www.umd.be. Finally, manual search was performed in the 250 other LSDBs reported for each of the 23 genes and are listed in Supp. Table S2.

To homogenize variant classification, Human Gene Mutation Database (HGMD) variant types have been matched as following: disease-causing mutation ("DM") = class 5, disease-causing

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: G. Collod-Béroud, "Genetics and Bioinformatics" research team, INSERM UMR_S910, Medical Genetics and Functional Genomics, Faculté de Médecine la Timone, 27 Bd Jean Moulin, 13385 Marseille Cedex 05, France. E-mail: Gwenaelle.collod-beroud@inserm.fr

Contract grant sponsors: Aix-Marseille Université, INSERM, European Union Seventh Framework Program (grant no. 305444).

mutation? ("DM?") = class 3, disease-associated polymorphism (DP) = class 1, functional polymorphism (FP) = class 2, and disease functional polymorphism (DFP) = class 2. When two or more databases had different variant classification, variant was classified as variant of unknown significance (VUS).

Reported frequencies from each variant from Exome Sequencing Project (ESP), Exome Aggregation Consortium (ExAC), 1000G, dbSNP (build 144) were extracted from the file provided by the Annovar Tool [Wang et al., 2010] as well as information from ClinVar (06-2015) (http://www.ncbi.nlm.nih.gov/clinvar/).

In silico predictions were performed with the UMD-Predictor tool [Salgado et al., 2016] through the corresponding Web service. Finally, data were merged into one table using a homemade Perl script.

What Are SVs and What Is the Importance of Reporting Them?

In 2013, the American College of Medical Genetics and Genomics (ACMG) recommended identification and return of SVs collected through NGS techniques such as WES and whole-genome sequencing (WGS) in diagnostic settings from a minimum set of 56 actionable genes as these variants, unrelated to the indication for which sequencing is ordered, are of medical value for patient care [Christenhusz et al., 2013; Green et al., 2013; ACMG Board of Directors, 2015]. These variants should be reported regardless of the age of the patient as preventive measures and/or treatment are available and individuals with pathogenic mutations might be asymptomatic for long periods of time. It was expected that the clinician would contextualize these variants for the patient in light of personal and family histories and physical examination.

Identification and reporting of these SVs led to a broad discussion in the last few years notably on: (1) clinicians' obligations or not to report them [Biesecker, 2013; Clayton et al., 2013; Gliwa and Berkman, 2013; van El et al., 2013], (2) patient's right "not to know" [Andorno, 2004; American College of Medical Genetics and Genomics, 2013; Scheuner et al., 2015], (3) extra workload needed for variant interpretation and confirmation [Dorschner et al., 2013; Hegde et al., 2015], (4) uncertain accuracy of genotypic predictions in the absence of familial segregation data [Burke et al., 2013], (5) or possibility of inadequate depth and breadth of sequencing coverage at clinically relevant locations [Park et al., 2015], but also (6) cost-effectiveness of this detection [Douglas et al., 2016], and finally (7) whether this effort would be compensated [Hegde et al., 2015]. One of the emerging questioning is our real capacity to label all true SVs.

How SVs Are Selected During the Variant Filtering Process?

The filtering of candidate variants by frequency in unselected individuals is a key step in any pipeline for the discovery of causal variants in Mendelian disease patients but also for the identification of SVs. Several databases are used to filter out polymorphisms (commonly variants with frequency above 1%). They can generally be assigned to the broad category of ***core*** (also named general or centralized) ***databases***. They are markedly different in terms of size, population diversity, and sequenced individual status (patient or obviously healthy) or enrichment for specific clinical conditions. It has also to be noted that many connections exist between them (Fig. 1), eliminating the need to consult multiple sources.

- **dbSNP**: The Single-Nucleotide Polymorphism database (dbSNP) (http://www.ncbi.nlm.nih.gov/snp) was established in September 1998, to address the need for a general catalog of genomic variation [Sherry et al., 2001]. dbSNP was initially composed of small-scale locus-specific submissions defined by flanking invariant sequences. Following the advent of high-throughput sequencing and the availability of complete genome assemblies for many organisms, dbSNP now receives a greater number of variants defined by sequence change at asserted locations on a reference sequence. dbSNP data evolved according to submissions from public laboratories and private organizations and now contains data from patients and controls of various ethnic groups. At present, dbSNP combines results from HapMap, 1000 Genomes, EVS, and ExAC projects (see below).
- **1000 Genomes Project**: 1000 Genomes (1000G) Project (http://www.1000genomes.org) includes today individual-level genotype data from 2,504 individuals from 26 populations [1000 Genomes Project Consortium et al., 2015]. Data are reconstructed genomes using a combination of low-coverage WGS, deep exome sequencing, and dense microarray genotyping. Populations are distributed as follows: 504 individuals with East Asian Ancestry, 489 with South Asian Ancestry, 661 with African Ancestry, 503 with European Ancestry, and 347 with American Ancestry. All these individuals are assumed to be healthy.
- **ESP**: Due to its goals, the NHLBI GO ESP (http://evs.gs.washington.edu/EVS/) contains in its last release (ESP6500SI-V2) exome variant data from 6,503 patients presenting with heart, lung, and blood disorders [Fu et al., 2013]. A subset of these data (ESP2500) having more stringent filtering criteria is available in the latest release of dbSNP (build 134) [Tennessen et al., 2012]. Samples are from unrelated individuals (samples showing first-degree to third-degree relatedness have been removed). Large-scale validation of the variants was not performed. However, sequencing validation of a small number of singletons (~200) and high-frequency SNP calls (~800) was performed [Tennessen et al., 2012]. The complete set of the SNP calls from the NHLBI ESP project is included in the dbSNP build-138.
- **ExAC:** ExAC (http://evs.gs.washington.edu/EVS/) aggregates and harmonizes exome sequencing data from 60,706 unrelated individuals sequenced as part of various disease-specific and population-genetic studies [Lek et al., 2015]. Individuals affected by severe pediatric diseases have been removed so this data set could serve as a reference set of allele frequencies for severe disease studies. ExAC contains today 7,404,909 high-quality variants, including 317,381 indels. Although 1000G and ESP are contributing projects, the majority of variants have very low frequency and 72% are absent from both 1000G and ESP [Lek et al., 2015].

The first step of filtering-out frequent variations is followed by matching the identified ACMG gene variants with the HGMD Professional release and ClinVar to identify variants known as causative. Origins of data and curation process are different for these two databases.

- **HGMD Professional**: The HGMD (http://www.hgmd.org) is a comprehensive collection of germline mutations in nuclear genes that underlie, or are associated with, human-inherited disease [Stenson et al., 2014]. HGMD is available in two versions: one public (permanently 3 years out of date and without any of the additional annotations) and one "Professional" obtainable by subscription (up to date version with curatorial comments). The mutation collection process is performed by automatic data mining systems that extract mutations from the various publication sources and checks their validity in comparison to reference

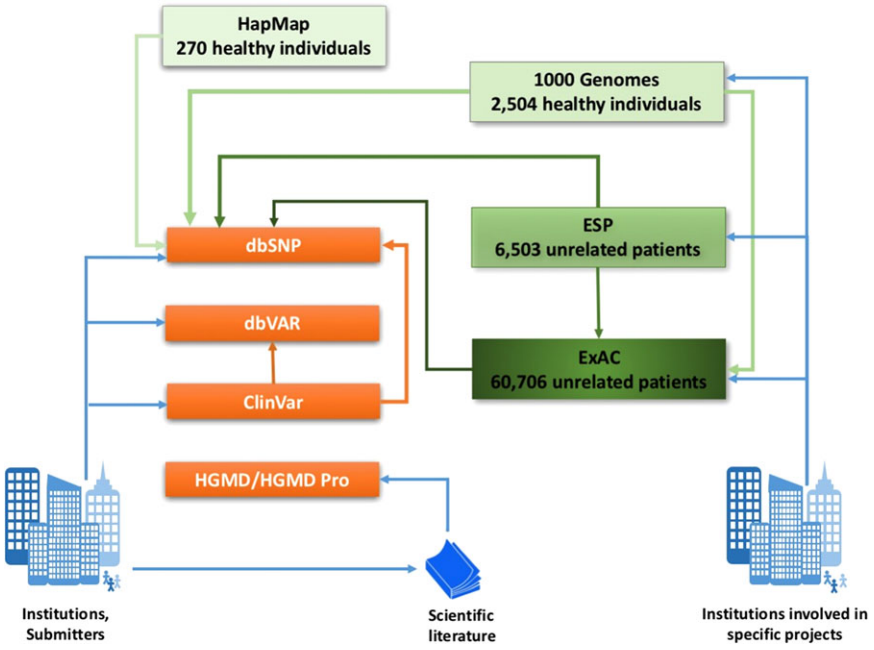

Figure 1. Interconnections between databases.

sequences using the international nomenclature. Manual curation is provided when necessary. By February 2016, the database contained over 127,000 different lesions detected in over 4,860 different genes in the public version and over 179,000 lesions in 7,189 different genes in the Professional version.

- **ClinVar**: ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/) is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence [Landrum et al., 2016]. ClinVar is seeded with records based on allelic variants described in OMIM (http://www.omim.org), GeneReviews or UniProt, variants submitted with clinical information to dbSNP, voluntary submissions from clinical testing laboratories, researchers, LSDBs, expert panels, and groups establishing professional guidelines. Submissions to ClinVar are categorized according to associated data as the type of submission (clinical testing, results part of research project, data extracted from the literature), the number of submitters, evidence that supports interpretation (genetic testing, family studies, comparison of tumor/normal tissue, animal models, etc.). ClinVar does not curate interpretations of clinical significance or arbitrate conflicts in interpretation. They invite the clinical genetics community to form expert panels, which should perform high-level curation for variant interpretations. ClinVar contains to date 173,216 records among which 85,642 have assertion criteria. Novel variants submitted to ClinVar are in turn submitted to dbSNP or dbVar.

Some teams also chose to evaluate pathogenicity of SVs according to in silico analyses. The most used and reliable prediction tools are: UMD-Predictor [Salgado et al., 2016], MutationTaster 2 [Schwarz et al., 2014], CADD [Kircher et al., 2014], Polyphen 2.2.2 [Adzhubei et al., 2013], SIFT 5.1.1 [Sim et al., 2012], Provean 1.1.3 [Choi et al., 2012], Mutation Assessor 2 [Reva et al., 2011], and CONDEL 1.5 [González-Pérez and López-Bigas, 2011] for missense variations and HSF [Desmet et al., 2009], ESE Finder [Smith et al., 2006], MaxEntScan [Yeo and Burge, 2004], and NNsplice [Reese et al., 1997] for variations potentially impacting splicing. Salgado et al. (2016) discusses these tools in this issue.

Can the Number of Individuals with Expected Actionable SVs Be Estimated?

This question is especially challenging as each identified variation is not linked to a specific sample for evident patient confidentiality. Various attempts have been made to evaluate the number of patients with SVs. These estimates are mainly based on variants already reported in HGMD Professional release followed by manual curation by specialists using PubMed and/or pathogenicity evaluation with different in silico tools that select only highly penetrant pathogenic mutations. If we restrict these different analyses to the ACMG recommended list of 56 genes, SVs have been found in a range from 1% to 5.6% of the participants (6/179 individuals [3.35%] [Xue et al., 2012], 19/1,000 participants [1.90%] [Dorschner et al., 2013], 12/1,092 participants [1.10%] [Olfson et al., 2015], 92/6,503 participants [1.41%] [Amendola et al., 2015], 623/11,068 participants [5.6%] [Gambin et al., 2015], 2/149 participants [2%] [Yavarna et al., 2015]).

Is the Number of Expected Actionable SVs Underestimated?

The reported range of 1%–5.6% of studied samples with SVs can be discussed. First, a high discordance among reviewers has been noticed by Amendola et al. (2015). Reviewers are likely to be inconsistent in their categorization and reports biased toward more pathogenic categories. Second, even if population minor allele frequency (MAF) is a useful factor for variant classification, data are also limited by population diversity and by the number of tested alleles. Some populations are poorly or not represented such as South Asian (Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka) and Latino individuals, or the Middle East population (Egypt, Iran, Turkey, Iraq, Saudi Arabia, Yemen, Syria,

United Arab Emirates, Israel, Jordan, Palestine, Lebanon, Oman, Kuwait, Qatar, Bahrain, and Cyprus). Databases might benefit from including a broader sampling of human diversity. Third, a possible overestimation of some SVs frequencies could be observed due to our inability to assess the MAF calculation. A bias could be introduced if MAF is based on a population enriched for pathogenic or likely pathogenic variants in specific ACMG genes. Several cohorts were, for example, enriched for lipid disorders, vascular disease, or chronic obstructive lung disease and are not a random sampling of the population. Finally, selection of variants according to their description in HGMD also introduces a bias as HGMD is not a comprehensive database. The impossibility nowadays to publish mutations in already known causing gene leads diagnostic laboratories to gather and store their variants in in-house databases. These data are submitted to core databases (as ClinVar) or to LSDBs only in the rare best case scenario.

In This Context, What About LSDBs for ACMG Genes?

LSDBs are a highly organized recording of variation data for specific genes. Lists of some LSDBs are available at the Human Genome Variation Society Website (HGVS, http://www.hgvs.org/locus-specific-mutation-databases), the Universal Mutation Database (UMD, http:www.umd.be), the Leiden Open Variation Database Website (LOVD, http://grenada.lumc.nl/LSDB_list/lsdbs), or the Gen2Phen Knowledge center (G2P, http://gen2phen.org/data/lsdbs). The majority of LSDBs presently available have been constructed with a small number of database management systems (DBMSs) among which the Leiden Open variation Database (LOVD, http:www.lovd.nl) [Fokkema et al., 2005] and the Universal Mutation Database (UMD, http:www.umd.be) [Béroud et al., 2005] that offer generic tools to build LSDBs. As they include data from a single gene, they collect all mutations and VUSs and often include unpublished data.

Numerous LSDBs are available for ACMG actionable genes (Supp. Table S2). All genes are represented in almost three different LSDBs, from which many involve the same DBMS (LOVD) but in different location (the number of variants was different). However, there are several LSDBs (18) that, although installed, have no variant documented.

LSDBs show a large heterogeneity in their contents and quality. Curation process varies largely among them. Highest quality in LSDB mutation collection process is provided by manual annotation of variants. This is a tedious but critical step since up to 10% of articles contain errors concerning mutation nomenclature: errors in type or position of mutations [Soussi et al., 2006] or use of a control sequence different from the current recognized reference.

Data commonly found in LSDBs are nucleotide position according to the reference sequence, exon number, description of the variation and nomenclature at the nucleotide (cDNA and genomic), and protein levels according to HGVS recommendations (http://www.hgvs.org/mutnomen/), reference of description (literature, diagnostic laboratories, etc.). For example:

(*FBN1*: sample IDXX c.3761G>A p.Cys1254Tyr g.48776092C>T [Stheneur et al., 2009] PMID19293843).

In some LSDBs, other data can be recorded such as associated disease, gender, transmission type (de novo, familial), geographic origin, specific location of the mutation at the protein level, consequences at the mRNA level, or experimental associated data. *In silico* analyses can also be available in some of them.

A wide heterogeneity is found for phenotypic data depending on the DBMS used. Phenotype description is usually reduced to single words in the great majority. Conversely, the UMD DBMS was developed notably to facilitate the collection of detailed phenotypes in view of performing genotype/phenotype correlation studies. Overall, the time spent to collect data and submit them according to LSDB needs is generally extensive and often restrains the involvement of large numbers of submitters and thus restrains their dissemination in the community.

Finally, LSDBs play a key role in the interpretation and classification of variants. It is widely accepted that classification of variation in genes is best performed by experts in those genes and/or pathology. Classification can be performed by individual curator(s) or an expert panel working with the curator and representing different areas of expertise (clinical, diagnostic, molecular, and computational). They display conclusion related to pathogenicity if a consensus has been reached. Pathogenicity was mainly based on familial segregation, evidence that supports a conclusion of pathogenicity, *in silico* prediction and frequency reported in core databases. For this, all these associated data have to be collected. Nevertheless, numerous LSDBs still do not provide manual annotation or classification of variants.

Table 1. Representation of the 318 Variations in Databases

Database	Number of variations found
LSDBs	**227**
LOVD Beacon	43
LOVD Share	96
UMD databases	123
HGMD Pro	**138**
Reported in HGMD and absent in LSDBs	**7**

Use Case

In order to face a real situation, we searched for lists of variations identified by exome sequencing in the 56 ACMG genes before any filtration by HGMD Pro. We based our analysis on lists published by Johnston et al. (2012). They performed exome sequencing on 572 participants selected for a range of atherosclerosis phenotypes, but not for personal or family histories of cancer. They analyzed nonsense, frameshift, splice-site, and nonsynonymous variants in 37 genes involved in cancer susceptibility syndromes among which 23 are part of the ACMG gene list. They provided a list of 451 variants among which 318 are carried by genes of the ACMG list. Reports and classification of each of these 318 variations were searched for by homemade Perl scripts. We queried "core" LOVD databases as LOVD share (http://databases.lovd.nl/shared/genes) and LOVD Beacon (http://mcupak.github.io/beacon-of-beacons/queries.html). UMD databases were queried online at http://umd.be. All other databases listed in Supp. Table S2 were also manually queried. Frequencies from ESP, ExAC, 1000G, and dbSNP (build 144), as well as *in silico* predictions with the UMD-Predictor tool [Salgado et al., 2016] were merged into one table (Supp. Table S1).

We first looked for the presence of variants in HGMD Pro (03/15/2016), LOVD Share, LOVD Beacon, and UMD databases. Results from all other databases listed in Supp. Table S2 (250 queried databases) were merged into a single category named "Other databases." Time to colligate all these data was estimated to be 16 hr. Of the 318 variations reported by Johnston et al. [2012], 138 (43.4%) were found in HGMD Pro (03/15/2016) (Table 1) and 227 (71.4%) in LSDBs. Representation in other databases was wide and only seven variations reported in HGMD (5%) were not found in LSDBs (Table 1). For the 180 variations not found in HGMD, 96 (53.3%)

Table 2. Representation of the 180 Variations Not Found in HGMD Pro (2016) Database

Database	Number of variations found
Not reported in LSDBs	**84**
Reported at least once in LSDBs	**96**
LOVD Beacon	23
LOVD Share	28
UMD databases	39

Table 3. Classification of Variants According to Databases

Database	Class 1	Class 2	Class 3	Class 4	Class 5	Total
HGMD Pro (03/15/2016)	15	9	81	0	**33**	138
HGMD Pro (2012)	11	7	51	0	**63**	132
Johnston classification (2012)	69	12	168	2	**8**	258
UMD classification	69	15	34	0	**5**	123
Other LSDBs classification	15	16	152	3	**6**	192

were at least reported in one LSDB and 84 (46.7%) were never reported, highlighting the added value of LSDB data (Table 2).

The most common of the cancer susceptibility syndromes analyzed was hereditary breast and ovarian cancer linked *to BRCA1/2* gene mutations with a combined frequency of ~1/500. Consequently, as Johnston et al. (2012), we considered that a variant with a MAF of $>1.5.10^{-2}$ was unlikely to cause a highly penetrant, rare, dominant disorder. Using ExAC allele frequencies (Supp. Table S1), a subset of 30/318 variations (9.4%) could be excluded with this criterion.

When variant classification is available in LSDBs, it usually follows recommendation in guidelines [Richards et al., 2015] with five gradations as (1) neutral variant, (2) likely neutral, (3) VUS, (4) likely causal, and (5) causal. This is not the case for HGMD Pro. To be able to compare variant classifications between all databases, HGMD annotations were matched to these five classes as following:

- DMs were matched with class 5 (causal);
- the annotation DM? corresponds to (1) variants initially classified as damaging in publications but with a degree of uncertainty, (2) variants reported by HGMD curators as having limited evidence for pathogenicity, and (3) variant for which pathogenicity was reconsidered after new evidence was provided. These variants were matched with class 3 (VUS);
- DPs are variants with evidence for a significant association with a disease/clinical phenotype along with additional evidence that the polymorphism is itself likely to be of functional relevance, although there may be no direct evidence of a functional effect. These variants were matched with class 1 (neutral);
- FPs correspond to variations that exert a direct functional effect but with no disease association reported as yet. These variants were matched with class 2 (likely neutral).
- DFPs correspond to variations that exert a direct functional effect, with no disease association reported as yet and displaying evidence of being of direct functional relevance. These variants were matched with class 2 (likely neutral).

When classification was conflicting between different LSDBs, class 3 (VUS) is assigned to variants.

In order to estimate the added value of LSDBs without involving another curation process, classification of variations not reported in databases were not evaluated.

Classifications of variants were compared between databases in order to identify the respective numbers of variants to be reported as SVs (Table 3). HGMD Pro (03/15/2016) reported 33 damaging variants (63 in 2012). Johnston et al. (2012) reported eight mutations after curation. Eleven variations in UMD databases and other LSDBs are described in class 5 (Table 4). In these 11 causal variants, five are not reported in HGMD Pro (Table 4), and three were classified as VUS by Johnston et al. (2012) (another one was not evaluated as described with poor quality, "class 0"). Three variations described as causal by Johnston et al. (2012) were not reported in databases (for two) or described as VUS (for one) (Table 4). In the 33 variants annotated as damaging by HGMD Pro 2016 (Table 4), 23 have been classified as nonpathogenic by LSDBs (class 1 to 3), three as "not reported," and six as causal (Table 5).

The proportion of SVs to report for cancer susceptibility syndromes in 572 exomes varies largely with 5.77% in HGMD Pro, 1.40% in Johnston's study, and 1.92% in LSDBs.

These results demonstrate the constant evolution of our knowledge leading to reannotation of variants in HGMD and in LSDBs over the years. Nevertheless, they also show that LSDBs give access to more information and help in classifying variants identified thanks to NGS.

Conclusion: How Can We Work Together?

LSDBs have evolved to serve many purposes to address the changing needs of the genetics community in evaluating and interpreting human genetic variation [Dalgleish, 2016]. There is no perfect generic design for LSDBs because of the heterogeneity of genetic diseases, associated phenotypes, and goals. Nevertheless, some recommendations have recently been published [Vihinen et al., 2016]. The more they offer phenotypic information, the less they are easy to maintain since quality of submitted data varies from center to center and over time. Another key challenge is to make the LSDB both easy to use and useful.

LSDBs are an ideal tool for integration and dissemination of data to the medical community. As expected, LSDBs contain more mutations than HGMD as they include up to 50% of unpublished variations (depending on the genes), often with phenotypic descriptions. Consequently, LSDBs are extremely useful tools, contributing to the identification of causative mutations, providing information about phenotypic patterns associated with a specific mutation, enabling researchers to define an optimal strategy for mutation detection, and helping in the characterization of SVs. LSDBs data could indeed significantly advance the interpretation of missense variants by facilitating estimates of the frequency of rare variants in patients presenting a given phenotype, of rare events co-occurring with pathogenic/nonpathogenic variants, of allele frequencies in specific populations and the association of variants with clinical or pathological features.

Today, data are still fragmented and various attempts have been made to develop unified databases. Nevertheless, they have mainly been unsuccessful, not only because of a lack of funding to create such databases. Indeed, first, LSDB have different goals reflected by different contents, infrastructures, and quality making them somehow hard to merge. Second, global efforts to gather genetic information from different databases and registries into a common global database have arisen [Bean and Hegde, 2016]. Such initiative must strongly benefit from LSDBS but this could be achieved only if they do not replace them, otherwise data sharing and expert curation will be compromised, especially as LSDBs are facing sustainability issues to offer accurate and updated data.

Database quality and accuracy depend on the involvement of all players from the data production chain. Data acquisition and

Table 4. Final List of Secondary Variants According to LSDBs

Gene	RefSeq AAChange	EVS6500	1000G_2015Aug	Exac03	UMD-predictor score	UMD-predictor prediction	UMD databases classification	Other LSDB classification	Number of reports in "Other databases"	HGMD Pro Classification (2012)	HGMD Pro Classification (03/15/2016)	Johnston's classification (2012)
BRCA1	NM_007294.3 **c.68_69del** p.Glu23Valfs*17	NA	NA	0,0002000	NA	NA	Not reported	5 - causal	1	5 - causal	Not reported	5 - causal
BRCA1	NM_007294.3 **c.547+2T>A**	NA	NA	NA	NA	NA*	Not reported	5 - causal	1	5 - causal	5 - causal	5 - causal
BRCA1	NM_007294.3 **c.688G>T** p.Glu230*	NA	NA	NA	100	Pathogenic	Not reported	5 - causal	1	Not reported	Not reported	0
BRCA2	NM_000059.3 **c.5946delT** p.Ser1982ArgfsX22	0.0002	NA	0,0003000	NA	NA	5 - causal	5 - causal	1	5 - causal	5 - causal	5 - causal
BRCA2	NM_000059.3 **c.8297delC** p.Thr2766AsnfsX11	NA	NA	NA	NA	NA	5 - causal	5 - causal	1	5 - causal	5 - causal	5 - causal
MUTYH	NM_001048171.1 **c.494A>G** p.Tyr165Cys	0.0022	0.000199681	0,0016000	90	Pathogenic	5 - causal	Not reported	0	**1-neutral**	Not reported	5 - causal
MUTYH	NM_001048171.1 **c.779G>A** p.Arg260Gln	0.0003	NA	0,0003000	84	Pathogenic	5 - causal	Not reported	0	5 - causal	Not reported	**3-VUS**
MUTYH	NM_001048171.1 **c.1145G>A** p.Gly382Asp	0.0038	0.00239617	0,0028000	90	Pathogenic	5 - causal	Not reported	0	**1-neutral**	Not reported	5 - causal
PTEN	NM_000314.4 **c.235G>A** p.Ala79Thr	NA	NA	0,0001000	87	Pathogenic	Not reported	4- probably causal	1	5 - causal	5 - causal	**3-VUS**
RET	NM_020630.4 **c.874G>A** p.Val292Met	NA	0.00379393	0,0006000	48	Polymorphism	Not reported	5 - causal	3	Not reported	5 - causal	**3-VUS**
SDHC	NM_003001.3 **c.43C>T** p.Arg15*	NA	NA	0,0000083	100	Pathogenic	Not reported	4- probably causal	2	5 - causal	5 - causal	5 - causal
Described as pathogenic only in Johnston et al. [2012]												
MUTYH	NM_001048171.1 **c.691C>T** p.Arg231Cys	NA	0.000199681	0,0000837	96	Pathogenic	**3-VUS**	Not reported	0	5 - causal	Not reported	4- probably causal
MUTYH	NM_001048171.1 **c.892-2A>G**	NA	0.00299521	0,0010000	NA	NA	Not reported	Not reported	0	5 - causal	Not reported	4- probably causal
BRCA2	NM_000059.3 **c.5482_5486del** p.Lys1828Valfs*4	NA	NA	NA	NA	NA	Not reported	Not reported	0	Not reported	Not reported	5 - causal

HSF prediction [Desmet et al., 2009]: alteration of the WT donor site, most probably affecting splicing.
Nucleotide numbering uses +1 as the A of the ATG translation initiation codon in the reference sequence, with the initiation codon as codon 1.

Table 5. Comparison of Classification of the 33 HGDM Causal Variants in LSDBs

Gene	RefSeq AAChange	EVS6500	1000G_2015Aug	Exac03	UMD-Predictor Score	UMD-Predictor Prediction	UMD database classification	Other LSDB classification	Number of reports in Other databases	Johnston classification	HGMD Pro 2012	HGMD Pro 03/15/2016
APC	NM_000038.4 c.607C>G p.Gln203Glu	0.0003	NA	0,0005000	47	Polymorphism	**3 - VUS**	**3 - VUS**	3	**3 - VUS**	Not Reported	5 - Causal
APC	NM_000038.4 c.3479C>A p.Thr1160Lys	0.0002	NA	0,0000995	93	Pathogenic	Not Reported	Not Reported	0	**3 - VUS**	5 - Causal	5 - Causal
APC	NM_000038.4 c.6821C>T p.Ala2274Val	0.0011	0.000199681	0,0010000	54	Probable polymorphism	**3 - VUS**	**3 - VUS**	1	**3 - VUS**	5 - Causal	5 - Causal
APC	NM_000038.5 c.7717A>G p.Ile2573Val	0.0003	NA	0,0001000	66	Probably pathogenic	Not Reported	Not Reported	0	**3 - VUS**	5 - Causal	5 - Causal
BRCA1	NM_007294.3 c.547+2T>A	NA	NA	NA	NA	NA	Not Reported	5 - Causal	1	5 - Causal	5 - Causal	5 - Causal
BRCA2	NM_000059.3 c.964A>C p.Lys322Gln	NA	0.000599042	0,0000582	5	Polymorphism	**3 - VUS**	Not Reported	1	**3 - VUS**	5 - Causal	5 - Causal
BRCA2	NM_000059.3 c.5946delT p.Ser1982ArgfsX22	0.0002	NA	0,0003000	NA	NA	5 - Causal	5 - Causal	1	5 - Causal	5 - Causal	5 - Causal
BRCA2	NM_000059.3 c.7504C>T p.Arg2502Cys	0.0012	0.000399361	0,0003000	47	Polymorphism	**3 - VUS**	Not Reported	9	**3 - VUS**	5 - Causal	5 - Causal
BRCA2	NM_000059.3 c.8297delC p.Thr2766AsnfsX11	NA	NA	NA	NA	NA	5 - Causal	5 - Causal	1	5 - Causal	5 - Causal	5 - Causal
MLH1	NM_000249.3 c.1742C>T p.Pro581Leu	NA	0.00119808	0,0001000	84	Pathogenic	Not Reported	**3 - VUS**	16	**3 - VUS**	5 - Causal	5 - Causal
MLH1	NM_000249.3 c.1963A>G p.Ile655Val	0.0032	0.00259585	0,0010000	29	Polymorphism	**1 - Neutral**	**3 - VUS**	26	**3 - VUS**	5 - Causal	5 - Causal
MLH1	NM_000249.3 c.1964T>C p.Ile655Thr	0.0002	NA	0,0000989	87	Pathogenic	**1 - Neutral**	**3 - VUS**	13	**3 - VUS**	5 - Causal	5 - Causal
MSH2	NM_000251.1 c.4G>A p.Ala2Thr	NA	NA	0,0004000	84	Pathogenic	**1 - Neutral**	1 - Neutral	2	**3 - VUS**	5 - Causal	5 - Causal
MSH2	NM_000251.1 c.815C>T p.Ala272Val	0.0004	NA	0,0002000	75	Pathogenic	**1 - Neutral**	**3 - VUS**	37	**3 - VUS**	5 - Causal	5 - Causal
MSH2	NM_000251.1 c.944G>T p.Gly315Val	NA	NA	0,0002000	96	Pathogenic	Not Reported	**3 - VUS**	2	**3 - VUS**	5 - Causal	5 - Causal
MSH2	NM_000251.1 c.1748A>G p.Asn583Ser	0.0002	NA	0,0000997	81	Pathogenic	**3 - VUS**	**3 - VUS**	6	**3 - VUS**	5 - Causal	5 - Causal
MSH2	NM_000251.1 c.1787A>G p.Asn596Ser	0.0002	NA	0,0003000	87	Pathogenic	**3 - VUS**	**3 - VUS**	24	**3 - VUS**	5 - Causal	5 - Causal
MSH2	NM_000251.1 c.2425G>A p.Glu809Lys	NA	0.000399361	0,0003000	54	Probable polymorphism	**3 - VUS**	**3 - VUS**	2	**3 - VUS**	Not Reported	5 - Causal
MSH6	NM_000179.2 c.1526T>C p.Val509Ala	0.0008	NA	0,0007000	50	Probable polymorphism	**2 - Likely neutral**	**3 - VUS**	7	**3 - VUS**	5 - Causal	5 - Causal
MUTYH	NM_001048171.1 c.74G>A p.Gly25Asp	NA	0.00179712	0,0011000	48	Polymorphism	**2 - Likely neutral**	**3 - VUS**	14	**3 - VUS**	5 - Causal	5 - Causal
MUTYH	NM_001048171.1 c.53C>T p.Pro18Leu	NA	0.00179712	0,0011000	60	Probable polymorphism	**2 - Likely neutral**	**3 - VUS**	14	**3 - VUS**	5 - Causal	5 - Causal
PTEN	NM_000314.4 c.235G>A p.Ala79Thr	NA	NA	0,0001000	87	Pathogenic	Not Reported	4- Probably causal	1	**3 - VUS**	5 - Causal	5 - Causal
RB1	NM_000321.2 c.411A>T p.Glu137Asp	0.0007	NA	0,0004000	63	Probable polymorphism	Not Reported	**3 - VUS**	6	**3 - VUS**	5 - Causal	5 - Causal
RB1	NM_000321.2 c.1966C>T p.Arg656Trp	0.0005	NA	0,0006000	93	Pathogenic	Not Reported	**3 - VUS**	3	**3 - VUS**	5 - Causal	5 - Causal
RET	NM_020630.4 c.785T>C p.Val262Ala	0.0002	0.000199681	0,0002000	63	Probable polymorphism	Not Reported	Not Reported	0	**3 - VUS**	5 - Causal	5 - Causal
RET	NM_020630.4 c.833C>A p.Thr278Asn	NA	0.00399361	0,0021000	57	Probable polymorphism	Not Reported	**3 - VUS**	1	**3 - VUS**	5 - Causal	5 - Causal
RET	NM_020630.4 c.874G>A p.Val292Met	NA	0.00379393	0,0006000	48	Polymorphism	Not Reported	5 - Causal	3	**3 - VUS**	Not Reported	5 - Causal
RET	NM_020630.4 c.1942G>A p.Val648Ile	7.7e-05	NA	0,0000908	48	Polymorphism	Not Reported	**3 - VUS**	4	**2 - Likely neutral**	5 - Causal	5 - Causal
SDHC	NM_003001.3 c.43C>T p.Arg15*	NA	NA	0,0000083	100	Pathogenic	Not Reported	4- Probably causal	2	5 - Causal	5 - Causal	5 - Causal
TSC1	NM_000368.4 c.1960C>G p.GLN654Glu	NA	0.00199681	0,0008000	69	Probably pathogenic	Not Reported	**3 - VUS**	11	**2 - Likely neutral**	5 - Causal	5 - Causal
TSC2	NM_000548.3 c.1939G>A p.Asp647Asn	7.7e-05	0.000399361	0,0004000	48	Polymorphism	Not Reported	**3 - VUS**	4	**3 - VUS**	5 - Causal	5 - Causal
TSC2	NM_000548.3 c.3430G>A p.Val1144Met	0.0002	NA	0,0002000	30	Polymorphism	Not Reported	**2 - Likely neutral**	2	**3 - VUS**	5 - Causal	5 - Causal
TSC2	NM_000548.3 c.5383C>T p.Arg1795Cys	0.0013	0.000798722	0,0012000	71	Probably pathogenic	Not Reported	**3 - VUS**	7	**2 - Likely neutral**	5 - Causal	5 - Causal

Nucleotide numbering uses +1 as the A of the ATG translation initiation codon in the reference sequence, with the initiation codon as codon 1.

enrichment rely mostly on diagnostic laboratories but they face two key obstacles:

1. *to receive complete information about patient's clinical presentation when a diagnostic test is ordered.* As such data are usually not available from the diagnostic laboratory itself, it is important to involve clinicians in this data sharing. Indeed, clinical descriptions are often scarce because of the lack of time in the course of the medical consultation. Clinicians are also insufficiently informed about the diagnostic laboratory possibility to transfer, in agreement with patient consent, accurate phenotypic data associated with mutations into databases.
2. *to justify the time spent on collecting data to their trustees.* This aspect is a real concern for diagnostic laboratories but also for clinicians. As previously mentioned, once a gene is described as disease-causing, most of the subsequent mutation identifications take place in a clinical setting. These data often present a low interest from journals because of the "lack of novelty." A large amount of curated sequence data therefore lies within the clinical laboratories for their own activity, waiting to be shared with the medical and research communities.

Attempts have been made to stimulate the sharing of those data by various mechanisms as microattribution, which unfortunately never expanded because of the lack of recognition by funding agencies or by trustees as a positive effort made by the investigators. However, although essential for optimum delivery of genetic healthcare and for medical research, the main difficulty for LSDBs is obtaining funding for the collection of such data. Only win-win approaches will be sustainable. The future may lie in public–private partnerships as illustrated by the successful BRCA-Share™ initiative [Béroud et al, 2016], to improve the detection of inherited risk of breast and ovarian cancers.

Acknowledgments

A.P. is supported by a PhD studentship from AFSMA (Association Française du Syndrome de Marfan et Apparentés). M.M. is supported by a PhD studentship from MENESR (Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche).

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. Nature 526:68–74.

ACMG Board of Directors. 2015. ACMG policy statement: updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing. Genet Med 17:68–69.

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet Chapter 7: Unit7.20.

Amendola LM, Dorschner MO, Robertson PD, Salama JS, Hart R, Shirts BH, Murray ML, Tokita MJ, Gallego CJ, Kim DS, Bennett JT, Crosslin DR, et al. 2015. Actionable exomic incidental findings in 6503 participants: challenges of variant classification. Genome Res 25:305–315.

American College of Medical Genetics and Genomics. 2013. Incidental findings in clinical genomics: a clarification. Genet Med 15:664–666.

Andorno R. 2004. The right not to know: an autonomy based approach. J Med Ethics 30:435–439; discussion 439–440.

Bean LJH, Hegde MR. 2016. Gene variant databases and sharing: creating a global genomic variant database for personalized medicine. Hum Mutat 37:559–563.

Béroud C, Hamroun D, Collod-Beroud G, Boileau C, Soussi T, Claustres M. 2005. UMD (Universal Mutation Database): 2005 update. Hum Mutat 26:184–191.

Beroud C, Letovsky SI, et al. 2016. BRCA Share: A Collection of Clinical BRCA Gene Variants. Hum Mutat 37.

Biesecker LG. 2013. Incidental variants are critical for genomics. Am J Hum Genet 92:648–651.

Burke W, Antommaria AHM, Bennett R, Botkin J, Clayton EW, Henderson GE, Holm IA, Jarvik GP, Khoury MJ, Knoppers BM, Press NA, Ross LF, et al. 2013. Recommendations for returning genomic incidental findings? We need to talk! Genet Med 15:854–859.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 7:e46688.

Christenhusz GM, Devriendt K, Dierickx K. 2013. Secondary variants–in defense of a more fitting term in the incidental findings debate. Eur J Hum Genet 21:1331–1334.

Clayton EW, Haga S, Kuszler P, Bane E, Shutske K, Burke W. 2013. Managing incidental genomic findings: legal obligations of clinicians. Genet Med 15:624–629.

Dalgleish R. 2016. LSDBs and how they have evolved. Hum Mutat 37:532–539.

Desmet F-O, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Béroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37:e67.

Dorschner MO, Amendola LM, Turner EH, Robertson PD, Shirts BH, Gallego CJ, Bennett RL, Jones KL, Tokita MJ, Bennett JT, Kim JH, Rosenthal EA, et al. 2013. Actionable, pathogenic incidental findings in 1,000 participants' exomes. Am J Hum Genet 93:631–640.

Douglas MP, Ladabaum U, Pletcher MJ, Marshall DA, Phillips KA. 2016. Economic evidence on identifying clinically actionable findings with whole-genome sequencing: a scoping review. Genet Med 18:111–116.

Fokkema IFAC, Dunnen den JT, Taschner PEM. 2005. LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat 26:63–68.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493:216–220.

Gambin T, Jhangiani SN, Below JE, Campbell IM, Wiszniewski W, Muzny DM, Staples J, Morrison AC, Bainbridge MN, Penney S, McGuire AL, Gibbs RA, et al. 2015. Secondary findings and carrier test frequencies in a large multiethnic sample. Genome Med 7:54.

Gliwa C, Berkman BE. 2013. Do researchers have an obligation to actively look for genetic incidental findings? Am J Bioeth 13:32–42.

González-Pérez A, López-Bigas N. 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88:440–449.

Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire AL, Nussbaum RL, O'Daniel JM, Ormond KE, Rehm HL, Watson MS, et al. 2013. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med 15:565–574.

Hegde M, Bale S, Bayrak-Toydemir P, Gibson J, Bone Jeng LJ, Joseph L, Laser J, Lubin IM, Miller CE, Ross LF, Rothberg PG, Tanner AK, et al. 2015. Reporting incidental findings in genomic scale clinical sequencing—a clinical laboratory perspective: a report of the Association for Molecular Pathology. J Mol Diagn 17:107–117.

Johnston JJ, Rubinstein WS, Facio FM, Ng D, Singh LN, Teer JK, Mullikin JC, Biesecker LG. 2012. Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. Am J Hum Genet 91:97–108.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310–315.

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, et al. 2016. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44:D862–D868.

Lek M, Karczewski K, Minikel E, Samocha K, Banks E. 2015. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291.

Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, Race V, Sistermans E, Sturm M, Weiss M, Yntema H, Bakker E, et al. 2016. Guidelines for diagnostic next-generation sequencing. Eur J Hum Genet 24:2–5.

Olfson E, Cottrell CE, Davidson NO, Gurnett CA, Heusel JW, Stitziel NO, Chen L-S, Hartz S, Nagarajan R, Saccone NL, Bierut LJ. 2015. Identification of medically actionable secondary findings in the 1000 Genomes. PLoS One 10:e0135193.

Park JY, Clark P, Londin E, Sponziello M, Kricka LJ, Fortina P. 2015. Clinical exome performance for reporting secondary genetic findings. Clin Chem 61:213–220.

Reese MG, Eeckman FH, Kulp D, Haussler D. 1997. Improved splice site detection in Genie. J Comput Biol 4:311–323.

Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 39:e118.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17:405–423.

Salgado D, Desvignes JP, Rai G, Blanchard A, Miltgen M, Pinard A, Lévy N, Collod-Beroud G, Béroud C. 2016. UMD-Predictor: a high-throughput sequencing compliant system for pathogenicity prediction of any human cDNA substitution. Hum Mutat 37:439–446.

Salgado D, Bellgard MI, Desvignes JP, Béroud C. 2016. How to identify pathogenic mutations among all those variations: variant annotation and filtration in the genome sequencing era. Hum Mutat.

Scheuner MT, Peredo J, Benkendorf J, Bowdish B, Feldman G, Fleisher L, Mulvihill JJ, Watson M, Herman GE, Evans J. 2015. Reporting genomic secondary findings: ACMG members weigh in. Genet Med 17:27–35.

Schwarz JM, Cooper DN, Schuelke M, Seelow D. 2014. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 11:361–362.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311.

Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 40:W452–W457.

Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ, Krainer AR. 2006. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum Mol Gen 15:2490–2508.

Soussi T, Ishioka C, Claustres M, Béroud C. 2006. Locus-specific mutation databases: pitfalls and good practice based on the p53 experience. Nat Rev Cancer 6: 83–90.

Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Gen 133:1–9.

Stheneur C, Collod-Béroud G, Faivre L, Buyck JF, Gouya L, Le Parc J-M, Moura B, Muti C, Grandchamp B, Sultan G, Claustres M, Aegerter P, et al. 2009. Identification of the minimal combination of clinical features in probands for efficient mutation detection in the FBN1 gene. Eur J Hum Genet 17:1121–1128.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337:64–69.

van El CG, Cornel MC, Borry P, Hastings RJ, Fellmann F, Hodgson SV, Howard HC, Cambon-Thomsen A, Knoppers BM, Meijers-Heijboer H, Scheffer H, Tranebjaerg L, et al. 2013. Whole-genome sequencing in health care. Recommendations of the European Society of Human Genetics. Eur J Hum Genet 21(Suppl 1):S1–S5.

Vihinen M, Hancock JM, Maglott DR, Landrum MJ, Schaafsma GCP, Taschner P. 2016. Human Variome Project quality assessment criteria for variation databases. Hum Mutat 37:549–558.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164.

Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper DN, Tyler-Smith C, 1000 Genomes Project Consortium. 2012. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am J Hum Genet 91:1022–1032.

Yavarna T, Al-Dewik N, Al-Mureikhi M, Ali R, Al-Mesaifri F, Mahmoud L, Shahbeck N, Lakhani S, AlMulla M, Nawaz Z, Vitazka P, Alkuraya FS, et al. 2015. High diagnostic yield of clinical exome sequencing in Middle Eastern patients with Mendelian disorders. Hum Gen 134:967–980.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11:377–394.